



# Prediction of hazards in coronary heart disease using decision trees

Priyadharshini S<sup>\*</sup>, Padmavathy S

Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, TamilNadu 636309, India

<sup>\*</sup>**Corresponding Author:** Faculty, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, TamilNadu 636309, India; Email – priya.maha.it@gmail.com

## Publication History

Received: 11 June 2015

Accepted: 19 July 2015

Published: 1 August 2015

## Citation


Priyadharshini S, Padmavathy S. Prediction of Hazards in Coronary Heart Disease Using Decision Trees. Discovery, 2015, 34(154), 48-53

## Publication License



© The Author(s) 2015. Open Access. This article is licensed under a [Creative Commons Attribution License 4.0 \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/).

## General Note

 Article is recommended to print as color digital version in recycled paper.

## ABSTRACT

In Human Life-Cycle process the cause of disease is a major problem which may cause morbidity and mortality in adults. The coronary heart disease (CHD) is one of the major problems that cause death to the human beings in this world. The cause of coronary heart disease is investigated based on their risk factors and it is related with the events from the past few decades by the medicinal fields. The risk factors investigated were: 1) clinical and 2) biochemical. Even though some statistical progress has been made in the diagnosis and treatment of coronary heart disease, further investigation is still needed in order to provide clear knowledge about the significant reduction of the coronary heart disease incidence. Here we develop the Data-Mining analysis system to get knowledge about the risk assessment factors. The objective of this study is to develop Data-Mining system for the assessment of heart-event related risk factors targeting in the reduction of the coronary heart disease hazards. A total of 298 cases were collected from the followings: 1) Cleveland clinic foundation 2) Hungarian institute of cardiology, budapet which may consists of large set of records with the several attribute list. Data mining analysis is carried out using Fuzzy ID3 (interactive dichotomizer 3) algorithms. Here the fuzzy rule based classification is one of the most popular approaches that generates a set of rules in order to generate a ID3 based decision tree which in turn extracts the event related risk factors that supports for the decision making process for the selection of medicinal therapy (i.e.) medical or surgical.

**Keywords:** Data mining, coronary heart disease, risk factors, Fuzzy ID3, decision trees, classification.

## 1. INTRODUCTION

Coronary heart disease (CHD) refers to the failure of coronary circulation to supply adequate circulation to cardiac muscle and surrounding tissue. It is the single most common cause of death in the developed countries, which is responsible for nearly two million deaths a year [1]. In coronary heart disease, the coronary arteries that supply heart muscle with oxygen and nutrients become narrowed by (plaque) in turn referred as atherosclerotic stenotic lesions which restricts the supply of nutrient rich blood and oxygen to the heart.

Extensive clinical and statistical studies had been identified several risk factors that may increase the hazards of CHD. The more risk factors one might have, the greater the risk of developing coronary heart disease. Also the greater the severity of each risk, the greater the overall risk. However this knowledge has not helped in the significant reduction of the CHD event. There are several risk factors that may contribute towards the development of coronary heart event. These risk factors are classified into two categories, clinical and bio-chemical. The first category is said to be act as a not modifiable one which cannot be recovered by any medicinal treatment, such risk factors identified are age, sex, family history of premature CHD, genetic attributes and operations. The second category is said to be act as a modifiable one which can be resolved by some medicinal treatment or surgical therapy, such risk factors are elevated cholesterol, smoking, hypertension, diabetes, high density lipoprotein, low density lipoprotein, triglycerides [2,3]. There are a number of other 'well-established' risk factors and protective factors that are also modifiable, but there are also numbers of other unknown factors that are not yet considered to be of great importance. The data which have been gathered gives information about the risk are complex and multifactorial thus making calculation of risk by just viewing the complete data as an extremely difficult task. Handling of such data is made possible by using data mining.

The objective of this study was to develop a data mining system based on decision trees for the assessment of CHD related risk factors targeting in the reduction of CHD hazards. Here the analysis in data mining is carried by using FUZZYID3 algorithm by means of some splitting criteria for extracting the rules.

## 2. BACKGROUND STUDY

The background studies are described one by one follows

### 2.1 Data Mining

Data mining is a process of extracting "hidden predictive information" from the large datasets. Here the events are said to be hidden in the large sets of datasets which consists of several risk factors of coronary heart disease of modifiable and not-modifiable events. The internal facts of datasets that contains different risk factors are associated and matched with the external facts of events to make decisional selection process of therapy.

### 2.2 Classification

Classification is one of the extensive advanced learning approaches used to identify the rules that divide the data into low, medium and high subgroups of the subjects.

### 2.3 Decision Trees

Decision trees are the reliable and effective decision making technique that provide high classification accuracy with a simple representation of gathered knowledge and they have been used in this medical decision making. It is one of the most widely used practical methods for the inductive references so as to make the right decision.

### 2.4 Id3

The divide-and-conquer approach to decision tree induction, sometimes called as top-down induction of the decision trees, was developed and refined by J.Ross Quinlan always been at the very fore front of decision tree induction. In turn the method that has been described using some of information gain criterion is essentially the same as one known as ID3. It then builds the decision tree from symbolic historic data so as to classify them with different values in turn decision tree chooses the attributes for decision making by referring the information gain values [4].

## 2.5 Fuzzy ID3

Fuzziness is incorporated into the ID3 algorithm at the node level by modifying the conventional decision function, with classical Shannon entropy, by the inclusion of different fuzzy measures. The fuzzy entropy considers a membership of a pattern to a class and helps to enhance the discriminative power of an attribute with some input feature values, Which is described in terms of some combination of overlapping membership values termed as low (L), medium (M) and high (H) as shown below[5].

$$F_i = [\mu_{\text{low}}(a_1)(F_i), \mu_{\text{medium}}(a_1)(F_i), \mu_{\text{high}}(a_1)(F_i) \dots \mu_{\text{high}}(a_n)(F_i)]$$

Where

$F_i = [a_1, a_2, \dots, a_n]$  implies set of attribute list.

$\mu_{\text{low}}$  = Membership value of low subjects.

$\mu_{\text{medium}}$  = Membership value of medium subjects.

$\mu_{\text{high}}$  = Membership value of high subjects.

The input values are divided into three partition which ranges from 0 to 1 in order to classify them. Here we define some of the threshold as a lower bound fraction of pattern that is allowed in an existing node.

## 3. RELATED WORKS

A previous study on the some dataset had showed that some of the risk factors identified are modified [6] therefore the risk of CHD of a patient is get reduced by means of proper control of these risk factors as it has been already published by EUROASPIRE studies [7]. The EUROASPIRE I, II surveys showed the high rates of modifiable coronary risk factors that acts as a preventive measures of coronary hazards [8]. Data mining system facilitates data exploration using data analysis methods with some sorts of frequent algorithms in order to discover un-known patterns. This system was also employed in several studies, where different system uses different algorithms for rule extraction and evolution, such systems are C4.5 decision trees [9], apriori algorithm [10], and k-means algorithm that all may focus only on the homogeneous group of data's that concentrates on low and high groups of classes.

## 4. MATERIALS AND METHODS

### 4.1 Data Collection

Data was collected from the different consecutives of the CHD subjects that were investigated in the past years by the supervision of some cardiologist. The dataset may consists of both the chemical and biochemical form of data such as age, sex, family history, triglycerides, smoking, operations, high density lipoprotein.

### 4.2 Data cleaning

The collected data may contain some of the noisy data, such as duplicated values, missing values. Here such fields are identified, extracted and filled. After data cleaning the collected cases are reduced, and this data are arranged in a particular manner to form a structured database.

### 4.3 Data Coding

The risk factors collected are coded with some defined threshold values to define the patient's subjective range. The criteria for data coding were provided by participating cardiologist and are as coded by American and European heart disease associations [11], the coded data are shown in Table 1.

**Table 1** Risk factors With Their Corresponding Coding

| Risk Factors | Code 1 | Code 2   | Code 3 |
|--------------|--------|----------|--------|
| Clinical     |        |          |        |
| Age          | 34-50  | 51-60    | 61-70  |
| Sex          | 0:Male | 1:Female |        |
| FH           | Y:Yes  | N:No     |        |
| HTN          | Y:Yes  | N:No     |        |

| Diabetes     | Y:Yes | N:no      |       |
|--------------|-------|-----------|-------|
| Bio Chemical |       |           |       |
| TC           | L<200 | M:201-240 | H>240 |
| HDL          | L<50  | M:50-60   | H>60  |
| LDL          | L<130 | M:130-150 | H>150 |
| TG           | L<120 | M:120-150 | H>150 |
| GLU          | N<110 | H>110     |       |

Table 1

L=>low; M=>Medium; H=>High

#### 4.4 Classification Using FUZZY Id3

The ID3 algorithm is the best known process for learning decision tree. Fuzzy based splitting criteria are used to focus on heterogeneous group of subjects in order to divide them as low, medium and high subgroup of classes [12]. The values are calculated by the following steps:

##### 4.4.1. Entropy (H)

The coded data values are used to predict the probability occurrences of the supporting cases and non supporting cases. The range is calculated using the formula shown below

$$\text{Info (D)} = - \sum_{i=1}^m p_i \log_2 (p_i)$$

Where

$p_i$  = probability (class i in dataset D);

m=number of class values.

##### 4.4.2 Information Gain (IG)

IG of an attribute A is used to select the best splitting criterion attribute. The highest info Gain is selected to build the decision tree.

$$\text{InfoGain(A)} = \text{Info(D)} - \text{Info}_A(\text{D})$$

Where A is the attribute investigated.

Here

$$\text{Info}_A(\text{D}) = \sum_{j=1}^V (|D_j|/|D|) \text{Info}(D_j)$$

Where

$|D_j|$  = number of observations with attribute value j  
in dataset D;

$|D|$ =total number of observations in dataset D;

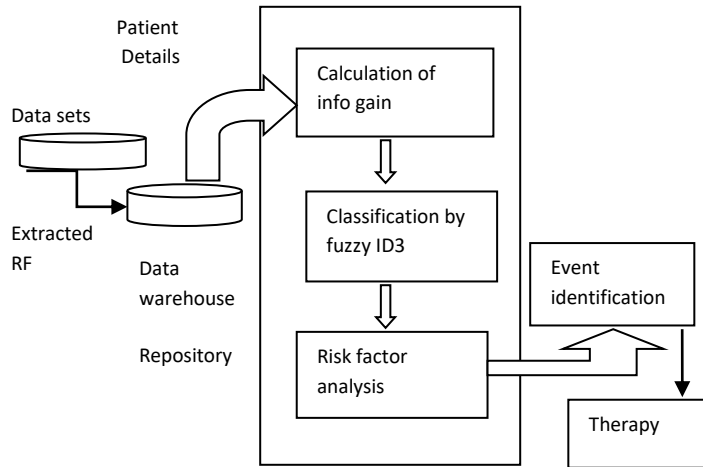
$D_j$  = sub dataset of D that contain attribute value j;

V=all attribute values.

## 5. PROPOSED STUDY

In this proposed study, we investigate how data mining based on decision trees can help for the evaluation of the risk of CHD and importance of each factor separately. The aim is to identify the most important risk factors based on the classification rules which enables for the better management of the patient targeting in the reduction of coronary hazards as well as the reduction of cost of therapy.

### 5.1. System Architecture



**Figure 1**

System Architecture for risk factor analysis

The (Figure 1) shown in the page explain us about the designed system architecture for risk factor analysis process. Here the extracted risk factors forms the data ware house repository in order to calculate the info gain values of the attributes which in turn is classified by the fuzzy ID3 so as to make the decision process.

### 5.3. Algorithm for the Proposed System

The algorithm of proposed system is defined by step by step following process:

Step 1: Create a root node "A".

Step 2: Select a Attribute list  $D = \{d_1, d_2, d_3...d_n\}$  from the database.

Step 3: Trace the values of  $d_1, d_2, d_3...d_n$  in the selected list Attribute list D.

Step 4: Match the attribute values of  $d_1, d_2, d_3...d_n$  with the defined classification rules of positive and negative range of values.

Step 5: Find amounts of positive rate and negative rate of the attribute values.

Step 6: Now calculate the Info of attribute value by using entropy formula.

Step 7: Define the subgroup of classes such as low, medium, and high with some classification rule.

Step 8: Trace the classes of where the attribute value lies and split it.

(i.e) if  $(D \in L \mid \mid D \in M \mid \mid D \in H)$

Step 9: Construct the Decision trees with the divided information values.

Step 10: The classes are then evaluated with the set of identified events.

Step 11: Finally the treatment is selected in according to the calculated D values.

### 5.4 Pattern Evaluation to Represent Knowledge

The following three different set of models by classifying a patient was investigated: (i) MI versus PCI or CABG, (ii) PCI versus MI or CABG and (iii) CAB versus MI or PCI For each of these models, the steps were carried out for data mining classification and pattern evaluation. Rules were extracted from different combinations of risk factors. A minimum of one to a maximum of ten risk factors were extracted from the different rules.

More specifically, selected rules were evaluated according to the importance of each rule. Each extracted rule was further evaluated by inspection of the number of cases from within the database that support the rule. Rules with a small number of records were ignored. We started with the strongest rules that mean the rules that were supported by most records in the database. We initially took the rules with one risk factor. In these rules the hierarchy of the risk of CHD evidently appeared i.e. the higher risk to the lower risk considering the number of cases. As second step we took the rules with two risk factors. Taking the risk factors with the highest percentage that we found in the first step, we checked which risk factor was the second higher for risk of CHD. The same strategy was used for the next step, taking into account rules with 3 risk factors, afterwards with 4 until 10 risk factors. This exercise finally concluded on the hierarchy of the risk factors.

## 5.5 Performance Measures

In order to evaluate our results we use following measures [13].

1. Correct classifications (%CC): it represents percentage of the correctly classified records that is  

$$\%CC = (TP + TN) / N$$
 Where  
 TP = correctly predicted positive rate  
 TN = correctly predicted negative rate  
 N = Total Numbers of records
2. False positive rate (%FP): it represents the wrongly predicted negative as positive one that is  

$$\%FP = FP / (TN + FP)$$
3. False negative rate (%FN): it represents the wrongly predicted positive as negative one that is  

$$\%FN = FN / (TP + FN)$$
4. Sensitivity: it is nothing a true positive rate  

$$\text{Sensitivity} = TP / (TP + FN)$$
5. Specificity: it is nothing a true negative rate  

$$\text{Specificity} = TN / (TN + FP)$$
6. Support: it tells about the number of cases which supports for the classified rule that is coverage.
7. Confidence: it tells about the number of cases that the rule can be applied so for, that is accuracy

## 6. CONCLUSION

In this study, a data mining system for the assessment of heart event related risk factors was carried out based on the fuzzy ID3 algorithm. Rules with risk factors like sex (male), smoking, high density lipoprotein, glucose, family history, and history of hypertension, were extracted. The modifiable risk factors can be monitored / lowered with the doctor's advice and medications so that the incidence of heart episodes can be lowered. It is anticipated that data mining could help in the identification of high, medium and low risk subgroups of patients, a decisive factor for the selection of therapy, i.e. medical or surgical. Moreover, the extracted rules could help to reduce CHD morbidity and possibly, mortality.

## REFERENCE

1. British Heart Foundation. (2008, Mar.8). European cardiovascular Disease statistics.
2. Euroaspire II study group, "lifestyle and risk factor and the use of drug the rapiesien coronary patients from 15 Countries", vol.22, 2002.
3. European society of cardiology-May 2008.
4. A comparative study of three Decision Tree algorithms: ID3, Fuzzy ID3 and Probabilistic Fuzzy ID3 by Guoxiu Liang at Informatics & Economics Erasmus University Rotterdam, the Neatherl and Augusts, 2005.
5. J. Han and M. Kamber, Data Mining, Concepts and Techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2001.
6. Assessment of the risk factors of coronary heart events bades on data mining with Decision Trees" by Minas A.karaolis IEEE, josephA. Moutiris and S. Pattichis, IEEE.
7. Fuzzy Decision Tree, Linguistic Rules and Fuzzy based knowledge network: Generation and Evalution by senior Member, IEEE, Kishori M.konwar, sushmita Mitra and, Sankar K.pal, fellow, IEEE.
8. M.Karaolis, J.A.Moutiris and S.pattichis, "Assessment of the risk of coronary heart event based on data mining", in proc. 8th IEEE Int.Conf, 2008.
9. T. Marshall, "Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values," Br.Med. Assoc. Public Health, vol. 8, p. 25, 2008.
10. Euroaspire study group," European society of cardiology survey of secondary prevention of coronary heart disease: principal results", Br.Med.Assoc public health, Vol.18,1997.
11. T.A.Pearson et al AHA guidelines of primary prevention of cardiovascular disease and stroke, Circulation 106(3) 2002.
12. P.N.Tan, Introduction to Data Mining: reading M.A: Addison-Wesley, 2006.
13. M. J. Zaki. "Mining non-redundant association rules". Data Mining and Knowledge Discovery Journal, pp.: 223-248, 2004.